The Performance and Agreement of Large Language Models on Radiology Resident in Training Examinations

The University of Texas at Austin
Dell Medical School

Mitul Gupta[1]; Christopher Kaufmann[1] MD, MS; Jack Virostko, PhD[1]
University of Texas Austin, Dell Medical School, Department of Diagnostic Medicine

Introduction

- While LLMs have demonstrated potential in diverse applications across medicine, there is limited and often unclear information on the accuracy and sources of publicly available AI tools such as ChatGPT-4, ChatGPT-3.5, Bard, and Claude specifically related to radiology domain knowledge.
- Radiology questions are designed to gauge not just simple factual knowledge but also require the integration of complex information provided through clinical scenarios to answer questions.
- For a radiologist to decide how, if at all, to utilize these assistive AI technologies, the output accuracy, relevance, and reliability needs to be better understood.
- ChatGPT has been tested across various knowledge such as the US Medical Licensing Examination, plastic surgery, urology, neurosurgery, and ophthalmology, where results varied from passing the USMLE to not performing as well on subspecialty specific examinations.
- There is a need for benchmarking in this space to understand the usefulness of models specifically for radiology. This study aims to establish a centralized benchmarking system to compare models and their performance over time on Radiology specific questions.

Methods

- The American College of Radiology Diagnostic In-Training Examination(DXIT) questions from 2020-2022 were utilized (n=172). Image-based questions were excluded due to the inability of LLMs to process images.
- The image dependent questions were distributed across various radiology disciplines including breast, cardiothoracic, GI/GU, musculoskeletal, neuroradiology, nuclear, pediatrics, ultrasound, interventional and radiology physics.
- Four publicly available AI LLMs platforms (Open AI GPT 3.5 & 4.0, Google Bard, and Claude) were used to evaluate for correct answer responses to the input test questions.
- Questions were queried from Nov 2023 to January 2024.
- One tailed, paired t-tests ($p<.05$) were used to compare overall model performance. Performance was also categorized by category.
- Post-hoc analysis was used to stratify question “difficulty” (<1 model with a question incorrect labeled “Easy”, 1-2 models incorrect labeled “Moderate”, >=3 models incorrect labeled “Hard”).
- For the “Hard” questions, an "agreement rate” was calculated. The maximum percentage of a given wrong answer across answer choices was taken as an indicator of model “agreement” on wrong answers.

Results

- Overall, GPT-4 performed best (76% ± 4.9%), followed by Google Bard (71% ± 4.3%), Claude (71% ± 1.2%), and GPT-3.5 (63% ± 6.9%) (Figure 1).
- GPT-4 was significantly better than Google Bard ($p = .003$) and GPT-3.5 ($p = .011$) with Google Bard performing significantly better than GPT-3.5 ($p = .033$)
- Performances across categories varied by model, with some performing better than others.
- Through the post-hoc stratification, 74 questions (43%) were “Easy”, 58 questions (34%) were “Moderate”, and 39 questions (24%) were “Hard” (Table 4). 1 question was answered “no answer” by most models and was excluded from this stratification.
- Performances of the models by stratified question difficulty category can be seen in Figure 3.

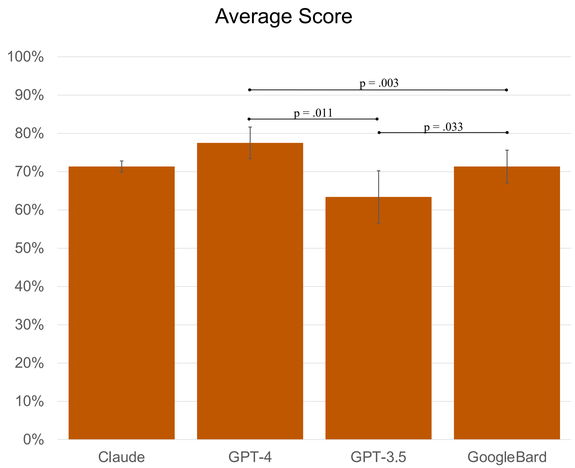

Figure 1. *Means and standard deviations of Large Language Model performance on DXIT questions*

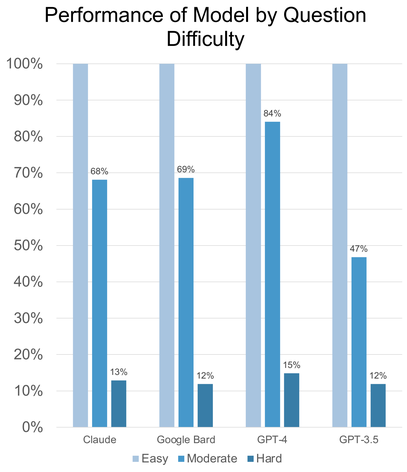

Figure 3. *Performance of the models on questions by stratified difficulty levels*

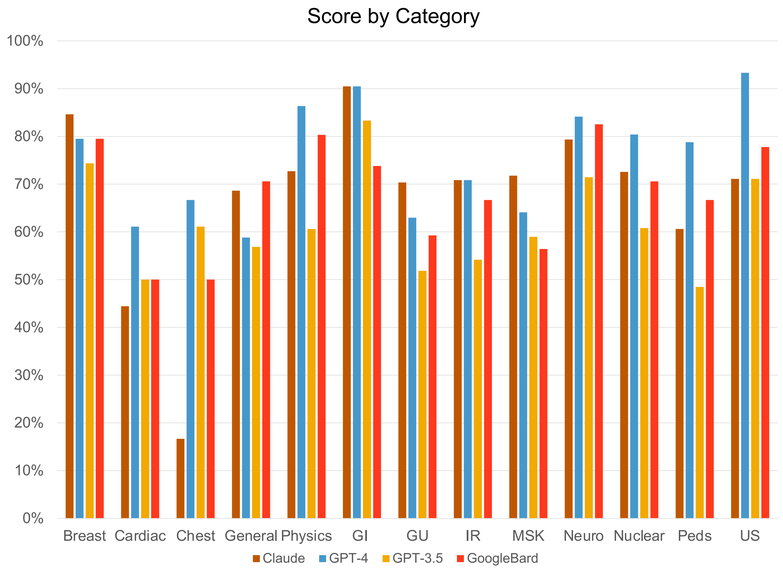

Figure 2. *Performance of the models across all tested categories*

Agreement

- On “Hard” questions, where 1 model or less correctly determined the answer, models agreed on the same (incorrect) answer 75% of the time, varying from 33% to 100% overall.
- Average agreement was 79%. 13/39 questions had a 100% agreement rate or concordance on the same wrong answer, with a generally right skewed distribution of agreement rate (Figure 4).

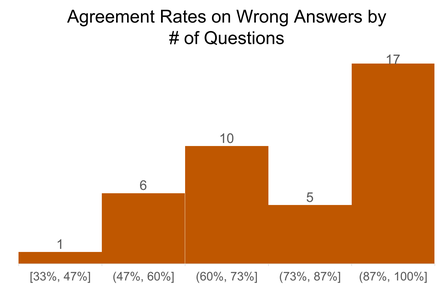

Figure 4. *Agreement of Models on wrong answers on “Hard” questions*

Conclusions

- The models all demonstrated performance at or near 70% on radiology domain specific questions, with variability between categories, some categories showing generally lower scores.
- Wrong answer analysis showed that models chose similar incorrect answers on “harder” questions, indicating a potential training bias or lack of training.
- As models are taught more information, standardized evaluation metrics on the performance, accuracy, and reliability of these models will need to be developed prior to integration in educational settings.

References

1. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K. et al. Large language models in medicine. Nat Med 29, 1930–1940 (2023). https://doi.org/10.1038/s41591-023-02448-8
2. Gilson, Aidan, et al. “How does CHATGPT perform on the United States Medical Licensing Examination? the implications of large language models for medical education and knowledge assessment.” JMIR Medical Education, vol. 9, 8 Feb. 2023, https://doi.org/10.2196/45312.
3. Huang, Yixing, et al. Benchmarking CHATGPT-4 on ACR Radiation Oncology in-Training (TXIT) Exam and Red Journal Gray Zone Cases: Potentials and Challenges for AI-Assisted Medical Education and Decision Making in Radiation Oncology, 2023, https://doi.org/10.2139/ssrn.4457218.
4. Bhayana, Rajesh, et al. “Performance of chatgpt on a radiology board-style examination: insights into current strengths and limitations.” Radiology, vol. 307, no. 5, 2023, https://doi.org/10.1148/radiol.230582.